Measuring inflectional and derivational complexity

Olga Sozinova[1] Christian Bentz[1,2] Tanja Samardžić[1]
olga.sozinova@uzh.ch chris@christianbentz.de tanja.samardzic@uzh.ch

[1]URPP Language and Space, University of Zürich
[2]DFG Center for Advanced Studies, University of Tübingen

Abstract

This paper presents experiments on measuring morphological complexity for the IWMLC 2019 shared task, which were performed on the Universal Dependencies data set (Track B). In particular, we are interested in differences between inflectional complexity and complexity of word formation processes, such as derivation and compounding. We propose an information-theoretic approach based on the unigram entropy of orthographic words, lemmas, and segments. We thereby aim to clarify how different types of inflection and word formation contribute to morphological complexity within a language, and how different morphological complexities are distributed across languages of the world.

1 Introduction

There is a panoply of studies proposing and testing morphological complexity measures (among others Juola 1998; Bane 2008; Liu and Xu 2011; Sagot and Walther 2011; Moscoso del Prado 2011; Ackerman and Malouf 2013; Baerman, Brown, and Corbett 2015; Kirjanov and Orekhov 2015; Bentz, Ruzsics, et al. 2016; Ehret and Szmrecsanyi 2016; Koplenig et al. 2017; Cotterell et al. 2018). However, quantitative measures often conflate different sources of morphological complexity, such as inflectional and derivational processes. Our aim is to tease these apart. To our knowledge, quantitative differences in inflectional and derivational morphology have so far been studied only within a psycholinguistic context (Milin et al. 2009).

We here propose to measure different dimensions of morphological complexity by calculating the unigram entropy of texts in three different versions: raw text, lemmatized text, and morphologically segmented text. We then compare unigram entropy values between the text versions, and evaluate how much information is carried by inflection versus other word formation processes.

This is a first step towards disentangling cross-linguistic differences in the usage of two principal morphological processes: inflection and word formation. In addition to cross-linguistic variation, we are interested in within-language ratios between these dimensions of complexity, since this is an important step to understand the nature of the morphological processes from a quantitative and usage-based perspective.

1.1 Inflection vs. word formation

We here follow the traditional distinction between inflection and word formation (e.g. Booij 2012). Inflection denotes processes that generate new word forms from a stem along certain grammatical categories, e.g. plural formation by suffixation (*dream* – *dreams*). Word formation, on the other hand, mainly refers to derivation and compounding. Derivation and compounding are processes that generate new words in the sense of a dictionary entry, e.g. generating an agent noun from a verb by adding an agentive suffix (*dream* – *dreamer*).

Languages differ with regards to the productivity of these morphological subdomains. For example, German has four nominal cases and Russian has six (Iggesen 2013), which might lead us to the assumption that Russian has more complex inflectional morphology (on nouns). Derivation, on the other hand, seems to be more productive in German. Some verbal derivatives in German are equivalent to Russian VPs, e.g. *durchschauen* 'see through' (derivated from *schauen* 'to watch') would be translated into Russian as *videt' naskvoz'*, which has the identical structure and meaning of the English translation. According to the backward dictionaries by Muthmann (2011, p. 527) and Zaliznjak (1980, p. 693), there are 53 derivatives from the German verb *sehen* 'to see', and only 7 derivatives from the Russian verb *videt'*.

Hence, derivational morphology seems richer in German.

However, such paradigmatic considerations of complexity do not take actual language production into account. Also, an overall morphological complexity measure would not reflect if and how these dimensions of complexity differ from each other.

2 Methods

2.1 Shannon Entropy

The gist of our approach is to measure differences in the Shannon entropy of unigrams (further defined below) between different conditions, i.e. different types of texts manipulated by automated morphological processing tools.

Shannon entropy is a measure of “uncertainty” or “choice”. The theoretical entropy of a set of symbols, given their probabilities $p_1, p_2, \dots p_n$ is defined as (Shannon and Weaver 1949, p. 50):

$$H(p_1, p_2, \dots p_n) = -K \sum_{i=1}^{n} p_i \ \log \ p_i, \tag{1}$$

where K is a positive constant determining the unit of measurement, and n is the number of different symbols. It is commonly assumed that $K = 1$, and the logarithm is taken to the base 2, which yields bits of information. In natural language texts, the symbols can be characters, combinations of characters, orthographic words, etc. Once we have chosen the set of symbols, and determined their probabilities, the “entropy measures how skewed the distribution is as a whole, that is, how deviant the most deviant member is, in addition to the second member, the third, and so on” (Hammarström and Borin 2011, p. 323).

In other words, the more skewed the probability distribution of symbols is towards high probabilities, the less information this distribution carries, and the less bits are needed for encoding it, reflected by lower entropy. Vice versa, if there are many unique symbols (and symbols of lower probability), then more bits are needed for encoding, there is more information, and higher Shannon entropy.

The main advantage of this metric is that it allows to objectively compare the differences in probability distributions of symbols in texts cross-linguistically. The probabilities of symbols are derived from corpora, and hence reflect language usage. However, we need to choose the symbols, and we should be aware that this choice has a crucial impact on the entropy measurements. Since the Universal Dependencies data set is tokenized by orthographic words, we choose these as the symbols to start with. However, in our analyses we deal not only with orthographic words, but also with their lemmas, as well as morphological segments. For simplicity, we will refer to all of these as *unigrams*.

The probability of unigrams is then calculated as their normalized frequency, i.e. the frequency of a particular unigram divided by the overall number of unigrams in the respective UD text. This is called the Maximum Likelihood (ML) approach. It is based on simplifying assumptions further discussed in Bentz, Alikaniotis, et al. (2017).

2.2 Text size dependence

One of the drawbacks of this method is dependency on the overall number of unigrams (tokens) given in a text. For example, the unigram entropy of orthographic words grows with the number of tokens. However, it stabilizes after ca. 50000 tokens (Bentz, Alikaniotis, et al. 2017). Hence, cross-linguistic comparisons are more meaningful for languages for which more than 50000 tokens are provided. In the UDtrack data set there are 26 languages with less than 50000 tokens. Below we provide token-increasing curves to illustrate the reliability of estimations.

Figure 1 shows how unigram entropy of the texts changes when the number of tokens (in this case orthographic words) increases. Russian and Czech corpora are given as examples here. In the figure we can see that after 50000 tokens, entropy growth slows down, meaning that the cross-linguistic difference between the entropies will also stabilize after this threshold.

3 Analyses of the UD data set

This section describes the core of our contribution. Here we provide the detailed technical steps in order to get the measures of inflectional and word formation complexities on the Universal Dependencies (UD) corpora.

Our measure of inflectional complexity is the difference between the unigram entropy of the raw text (henceforth H_raw) and the unigram entropy of the lemmatized text (henceforth H_lemmas). Our measure of word for-

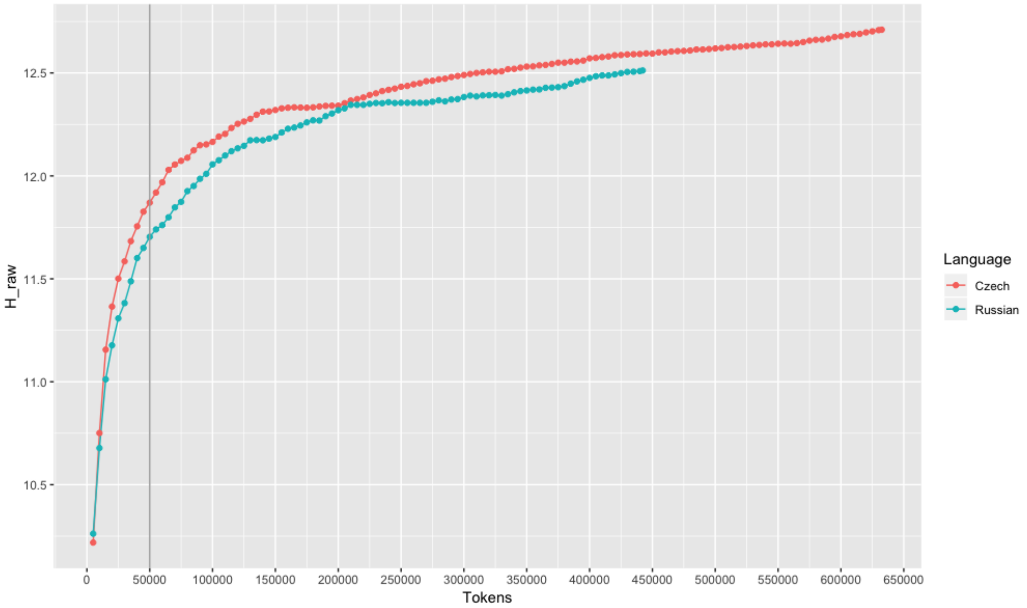

Figure 1. Entropy stabilization with increasing number of tokens (in this case orthographic words).

mation complexity, on the other hand, is the difference between the unigram entropy of the lemmatized text and the unigram entropy of the morphologically segmented text (henceforth H_segments). The first measure is denoted in the result CSV table as SBS_INF, and the second measure is denoted as SBS_DER.

For our calculations we use the second and the third columns of the CoNLL-U files, i.e. orthographic word tokens and lemmas. There are hence three separate calculations of unigram entropies for each UD text:

- Firstly, we calculate unigram entropy of the original raw texts (column of orthographic word tokens) from the UDtrack data set (i.e. H_raw).
- Secondly, we do the same for the lemmatized text, given in the second column of the CoNLL-U format (lemmas column) (i.e. H_lemmas).
- Thirdly, we apply morphological segmentation (further detailed below) on the lemmatized texts, and then calculate the unigram entropy of morphological segments (i.e. H_segments).

For morphological segmentation, we use Morfessor 2.0 (Virpioja et al. 2013), which is a state-of-the-art tool for unsupervised morpheme segmentation. We use the default settings (baseline model, Viterbi segmenting algorithm). The resulting texts contain segments recognized by Morfessor. Note that Morfessor 2.0 was trained and tested mainly on English and Finnish, and works well for concatenative morphology. The results for languages with non-concatenative word formation, e.g. Arabic and Hebrew, are probably less meaningful. In our approach, any subword – more precisely "sublemma" – patterns found by Morfessor function as an approximation of derivational morphemes and compounding.

For illustration, the following example shows our steps in processing the corpora, and the resulting unigram entropy calculations on a toy text (Table 1):

Table 1. Raw, lemmatized and segmented text

Raw	Lemmatized	Segmented
Roses are red	Rose be red	Rose be red
Plums are blue	Plum be blue	Plum be blue
Sugar is not red	Sugar be not red	Sugar be not red
But sugarplums are blue	But sugarplum be blue	But sugar plum be blue

Firstly, we calculate the Shannon entropy (Equation 1) of unigrams for the orthographic words. The probability p_i of the i^{th} unigram is calculated as a relative frequency of the unigram in a given text (ML method described above). In order to calculate the unigram entropy of the raw text, we have to create a frequency dictionary, and count relative frequencies for each unigram. In the toy text, $n = 10$, i.e. there are ten unique unigrams (types), whereas the number of unigram tokens is 14, since some types occur several times. The frequency dictionary looks as follows: {'roses': 1, 'are': 3, 'red': 2, 'plums': 1, 'blue': 2, 'sugar': 1, 'is': 1, 'not': 1, 'but': 1, 'sugarplums': 1}.

Equation 2 shows the steps of calculation which lead to the resulting entropy value of 3.182 bits.

$$\begin{aligned} H\_raw &= -(p(\text{roses})\log_2 p(\text{roses}) + p(\text{are})\log_2 p(\text{are}) + ... \\ &+ p(\text{sugarplums})\log_2 p(\text{sugarplums})) \\ &= -(\frac{1}{14}\log_2\frac{1}{14} + \frac{3}{14}\log_2\frac{3}{14} + ... + \frac{1}{14}\log_2\frac{1}{14}) = 3.182 \end{aligned} \tag{2}$$

After lemmatization and segmentation the frequency dictionaries and number of tokens change. For example, the unigrams ‘are’ and ‘is’ are now collapsed to ‘be’, which then has a frequency of five. This particular change in frequencies decreases unigram entropy. In this example, $H\_lemmas = 2.95$, and $H\_segments = 2.84$.

Inflectional and derivational complexities are then calculated as unigram entropy differences between the three conditions:

$$\begin{aligned} SBS\_INF &= H\_raw - H\_lemmas = 3.182 - 2.95 = 0.232 \\ SBS\_DER &= H\_lemmas - H\_segments = 2.95 - 2.84 = 0.11 \end{aligned} \tag{3}$$

The obtained values show that the example text has higher inflectional complexity than word formation complexity. In fact, seven tokens can be lemmatized (‘roses’, ‘are’, ‘plums’, ‘are’, ‘is’, ‘sugarplums’, ‘are’), while only one token is split into segments (‘sugarplums’).

4 Results and discussion

The results table Sozinova.csv contains values for the measures SBS_INF and SBS_DER calculated for all 63 texts given in the UDtrack data set.

Figures 2 and 3 depict the distribution of SBS_INF and SBS_DER across different language groups. Note that two outliers (Uyghur with an average of 4.47 for SBS_INF, and Korean with an average of 1.85 for SBS_DER) are not included in these plots; see the discussion on these values in the Section 4.2.

In the Figures 2 and 3 we can see that language groups tend to cluster around specific values; this is true, for example, for Slavic, Germanic, Romance and Finnic languages.

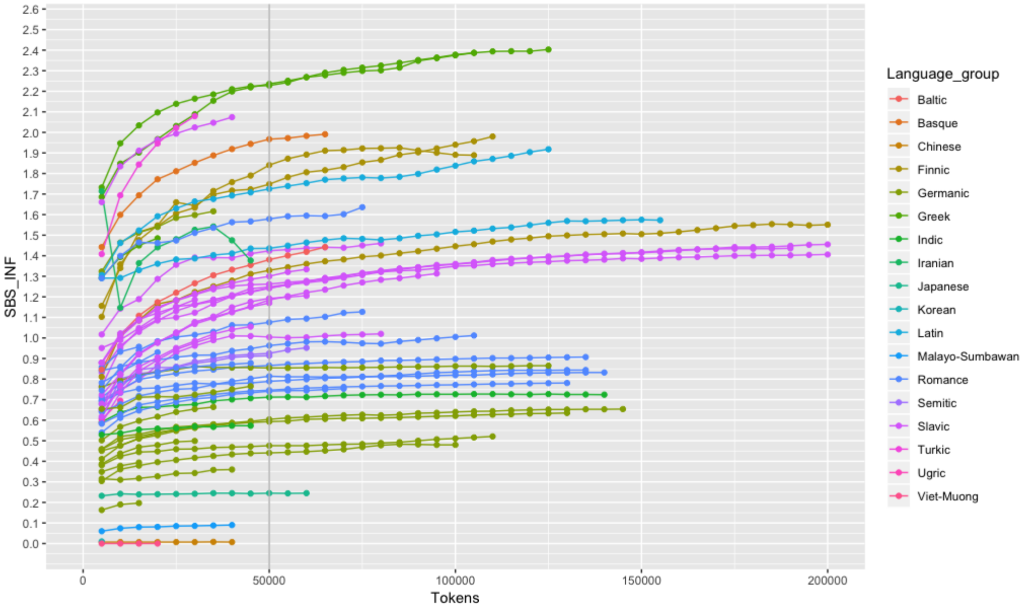

Figure 2. Distribution of SBS_INF

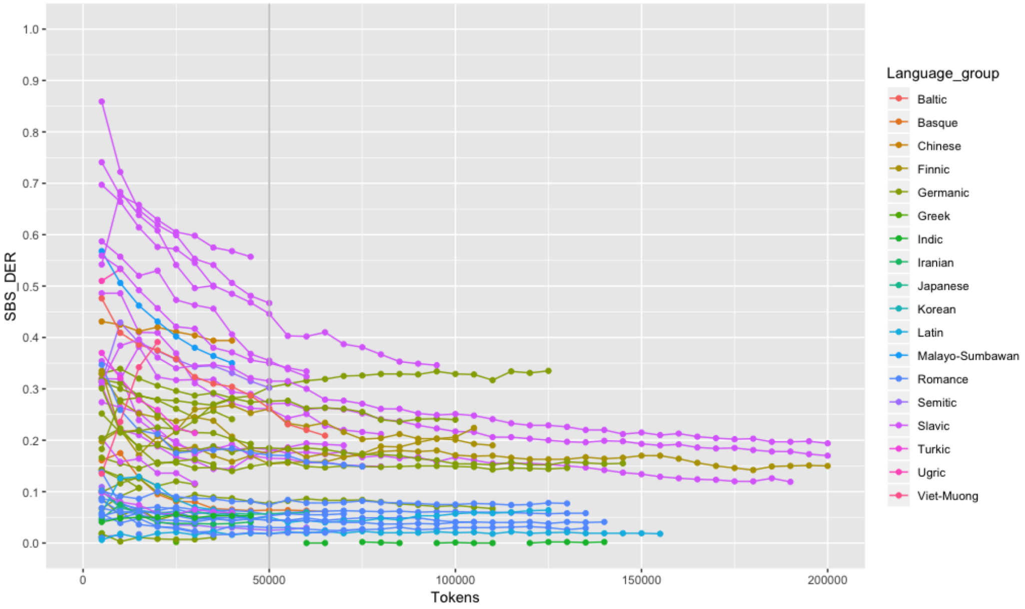

Figure 3. Distribution of SBS_DER

Below we discuss in more detail the case of Germanic and Slavic languages against the backdrop of German and Russian morphological complexities as hinted at earlier. In Section 2, we gave some intuitions on how inflectional and derivational complexity might differ between German and Russian. Below we plot the values for SBS_INF and SBS_DER (Figures 4 and 5) and can compare all Slavic and Germanic languages in the sample (excluding dead languages, i.e. Old Church Slavonic and Gothic). Given the bias that occurs due to the number of tokens, we performed additional calculations on different portions of texts (50000, 100000, etc. tokens) and compared the values with those around the point of 50000 tokens (Figures 4, 5).

4.1 Slavic vs. Germanic

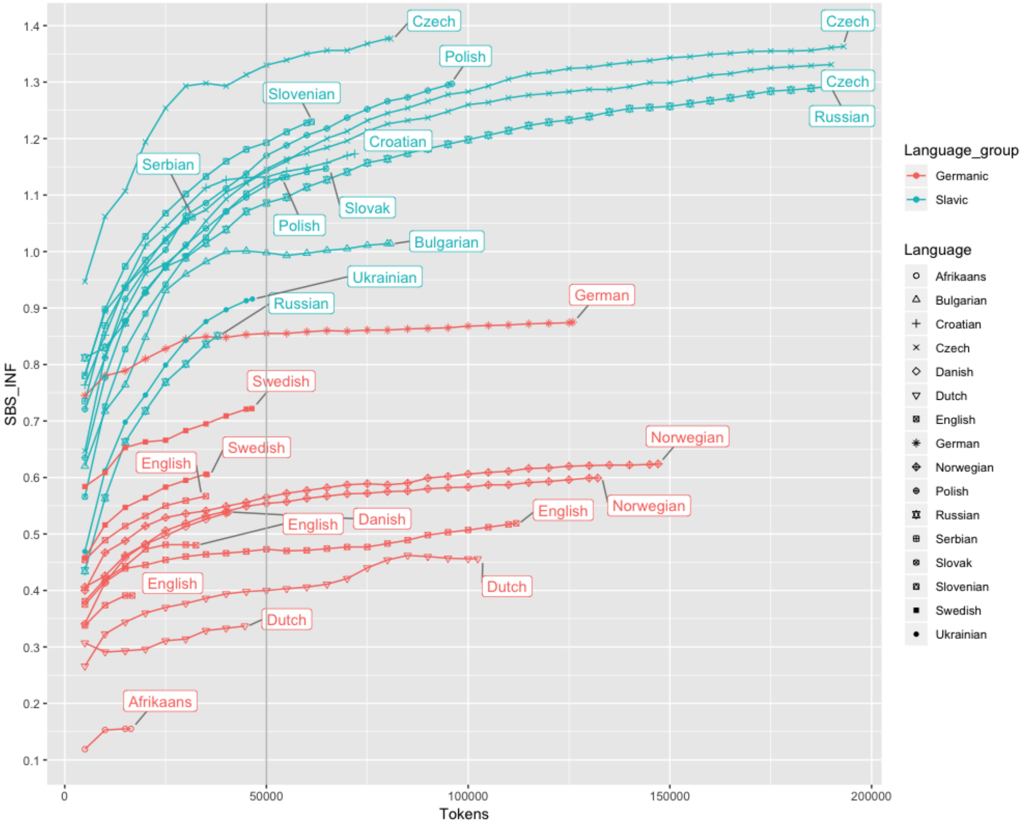

Figure 4. Slavic vs. Germanic: inflection

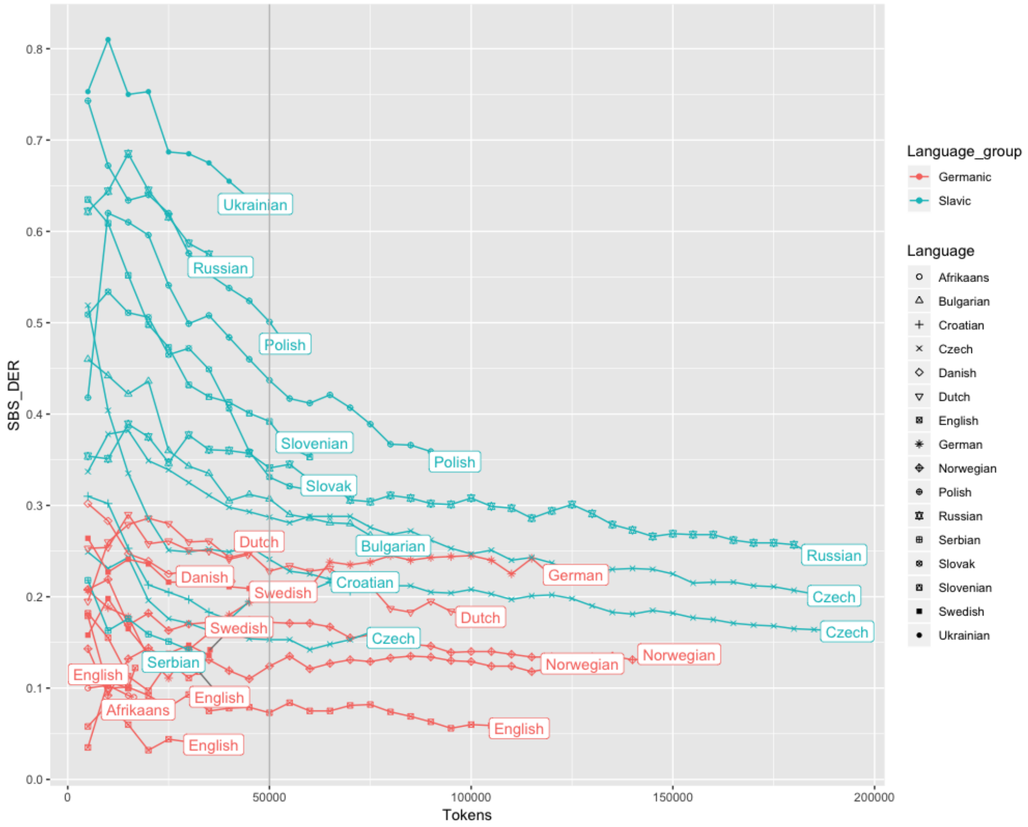

Figure 5. Slavic vs. Germanic: derivation

In the Figures 4 and 5 we can see that (according to our method) Russian has more complex inflection *and* word formation processes than German, which is unexpected given the discussion above. In fact, this trend seems to hold for Slavic versus Germanic languages in general.

There are some noticeable exceptions though. For example, Croatian and Serbian have high values for inflection (among the other Slavic languages), but low values for derivation (rather among the Germanic languages). This effect might occur due to the different orthographic rules in comparison to the other Slavic languages. Croatian and Serbian writing systems are phonetic, not phonemic as in other Slavic languages. For example, in Serbian, the words ***Srb***-*ija* 'Serbia' and ***srp***-*ski* 'Serbian' have different roots due to regressive assimilation. In Russian, on the other hand, both words have the same root: ***Serb***-*ia* and ***serb***-*skij*.

As a consequence of this orthographic difference, Croatian and Serbian segments are more varied than in the case of other Slavic languages. This leads to a relatively small difference in entropy between lemmatized and segmented text. In other words, Serbian and Croatian text remain complex after segmentation.

4.2 Some Problems and Caveats

Our results here depend on the respective orthographic words and lemmas given in the UD data set, as well as the performance of the software Morfessor. Some problems and caveats with this approach are discussed below.

Some of the extreme values represented in our measures are very likely artefacts of our methods. SBS_INF ranges from -0.037 for the Korean-Kaist corpus to 4.735 for Uyghur. SBS_DER ranges from -0.057 for Greek to 2.082 for the Korean-GSD corpus.

For both Korean UD texts, the lemma entropy is slightly higher than the raw text word entropy, such that the difference between raw and lemma entropy is negative (-0.003 and -0.037 respectively). This is a highly unexpected result from a theoretical point of view. Lemmatization should always decrease entropy, since inflectional information is neutralized. In the case of the Korean UD texts, the changelog in README.md for Korean-GSD says that lemmas are added using the KOMA morphological analyzer. It is described in Lee and Rim (2005) as well as Lee and Rim (2009). The description there reveals that what is given as "lemma" in the Korean UD is actually a morpheme segmentation of the original orthographic word (called Eojeol). Hence, rather than being neutralized, the inflectional morphemes are delimited by '+'. For example, *hag-gyo-e* 'to school' is written as one orthographic word in the Korean script Hankul, and the morphological analysis by KOMA is *hag-gyo+e*, since 'e' is the inflectional dative case marker indicating destination.

For our analyses of Korean UD texts this means that the numbers for inflectional and derivational complexity do not give a realistic picture, since what is given as lemmas in the UD texts are actually just the original orthographic words with segmentation markers added. This certainly explains why the raw unigram entropy is even lower than the lemma entropy.

According to our results, Uyghur has the highest inflectional complexity of all the languages in the sample, with a raw text to lemma entropy difference of 4.735 bits per unigram. However, the lemma entropy to segment entropy difference is -0.045, meaning that further segmentation actually increases, rather than decreases entropy. According to the descriptions in the Grammar of Uyghur by Tömür (2003), finding an exceptionally high inflectional complexity makes sense. Verbs are inflected for voice, negation, tense, mood, etc. and nouns are inflected for number, and overall 10 cases (Tömür 2003, pp. 34–37). However, Uyghur is also reported to have rich derivational processes (see an example in Tömür 2003, p. 29). Thus, it is likely that there is some artefact in the segmentation results obtained by using Morfessor.

Uyghur is written in Arabic script, and the problem potentially lies there, since Morfessor was not trained on non-latin scripts. This could also explain why Greek and Ancient Greek have negative values for the lemma to segment entropy difference. In Greek there are definitely productive derivational prefixes in the lemmas, which should be – but are not – further segmented by Morfessor. These issues need to be further explored.

5 Conclusion

In this paper, we presented a new approach to tease apart different dimensions of morphological complexity: inflection and word formation. We propose to compare unigram entropies calculated on the raw, lemmatized and segmented versions of the UD texts. This is an information-theoretic, usage-based and reproducible means of understanding inflectional and derivational complexities. We have encountered several obstacles during our investigation, for instance, unexpected negative values for our proposed complexity measures. These are artefacts deriving from certain data processing decisions taken in the UD, in conjunction with our methods. However, there are potential remedies for these to be implemented in future research.

References

Ackerman, Farrell and Robert Malouf (2013). “Morphological organization: The low conditional entropy conjecture”. In: *Language* 89.3, pp. 429–464.

Baerman, Matthew, Dunstan Brown, and Greville G Corbett, eds. (2015). *Understanding and measuring morphological complexity.* Oxford University Press.

Bane, Max (2008). “Quantifying and measuring morphological complexity”. In: *Proceedings of the 26th west coast conference on formal linguistics*. Cascadilla Proceedings Project Somerville, MA, pp. 69–76.

Bentz, Christian, Dimitrios Alikaniotis, Tanja Samardžić, and Paula Buttery (2017). “Variation in word frequency distributions: Definitions, measures and implications for a corpus-based language typology”. In: *Journal of Quantitative Linguistics* 24.2-3, pp. 128–162.

Bentz, Christian, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardzic (2016). “A comparison between morphological complexity measures: typological data vs. language corpora”. In: *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pp. 142–153.

Booij, Geert (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

Cotterell, Ryan, Christo Kirov, Mans Hulden, and Jason Eisner (2018). “On the Complexity and Typology of Inflectional Morphological Systems”. In: *arXiv preprint arXiv:1807.02747*.

Ehret, Katharina and Benedikt Szmrecsanyi (2016). “An information-theoretic approach to assess linguistic complexity”. In: *Complexity, isolation and variation*. Ed. by Raffaela Baechler and Guido Seiler. Berlin: de Gruyter.

Hammarström, Harald and Lars Borin (2011). “Unsupervised learning of morphology”. In: *Computational Linguistics* 37.2, pp. 309–350.

Iggesen, Oliver A. (2013). “Number of Cases”. In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: `https://wals.info/chapter/49`.

Juola, Patrick (1998). “Measuring linguistic complexity: The morphological tier”. In: *Journal of Quantitative Linguistics* 5.3, pp. 206–213.

Kirjanov, Denis and Boris Orekhov (2015). “Complex networks-based approach to transcategoriality in the Bashkir”. In: *Proceedings of the New Developments in the Quantitative Study of Languages*, p. 34.

Koplenig, Alexander, Peter Meyer, Sascha Wolfer, and Carolin Müller-Spitzer (2017). “The statistical trade-off between word order and word structure–Large-scale evidence for the principle of least effort”. In: *PloS ONE* 12.3, e0173614.

Lee, Do-Gil and Hae-Chang Rim (2005). “Probabilistic models for Korean morphological analysis”. In: *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Lee, Do-Gil and Hae-Chang Rim (2009). "Probabilistic modeling of Korean morphology". In: *IEEE Transactions on Audio, Speech, and Language Processing* 17.5, pp. 945–955.

Liu, Haitao and Chunshan Xu (2011). "Can syntactic networks indicate morphological complexity of a language?" In: *EPL (Europhysics Letters)* 93.2, p. 28005.

Milin, Petar, Victor Kuperman, Aleksandar Kostic, and R Harald Baayen (2009). "Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation". In: *Analogy in grammar: Form and acquisition.* Ed. by James P. Blevins and Juliette Blevins. Oxford: Oxford University Press, pp. 214–252.

Moscoso del Prado, F (2011). "The Mirage of morphological complexity". In: *Proc. of the 33rd Annual Conference of the Cognitive Science Society*, pp. 3524–3529.

Muthmann, Gustav (2011). *Rückläufiges deutsches Wörterbuch: Handbuch der Wortausgänge im Deutschen, mit Beachtung der Wort- und Lautstruktur.* Reihe Germanistische Linguistik. De Gruyter. ISBN: 9783110920666. URL: `https://books.google.ch/books?id=JEUjGoY1QgwC`.

Sagot, Benoıt and Géraldine Walther (2011). "Non-canonical inflection: data, formalisation and complexity measures". In: *International Workshop on Systems and Frameworks for Computational Morphology.* Springer, pp. 23–45.

Shannon, Claude E and Warren Weaver (1949). "The mathematical theory of information". In:

Tömür, Khāmit (2003). *Modern Uyghur grammar: morphology.* Vol. 3. Yıldız.

Virpioja, Sami, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo (2013). *Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline.* D4 Julkaistu kehittämis- tai tutkimusraportti tai -selvitys, p. 38. URL: `http://urn.fi/URN:ISBN:978-952-60-5501-5`.

Zaliznjak, Andrej Anatol'evich (1980). *Grammatical dictionary of Russian language [Grammaticheskij slovar' russkogo jazyka].* Russkij jazyk.